VRI: 3D QSAR at Variable Resolution

GORDON M. CRIPPEN
College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109-1065

Received 19 February 1999; accepted 16 June 1999

ABSTRACT: VRI (Variable Resolution Invariants) is a new approach to quantitative structure–activity relations that makes use of three-dimensional features of molecules at different levels of spatial resolution as well as levels of resolution in atomic properties. These descriptors are independent of any numbering of the atoms of a molecule. They are also independent of rigid translation and rotation of a given conformer, which avoids problems with aligning different molecules or docking them with a receptor site model. Steric effects, stereospecificity, substituent effects, lipophilicity, and conformational flexibility are all dealt with in a single, natural formalism. Simple datasets can be fitted using a small number of descriptors corresponding to low-resolution descriptions of the molecules. More complicated data can require additional descriptors that recognize finer details of three-dimensional structure and physico-chemical properties. Overfitting due to the large number of descriptors is handled by partial least-squares analysis with crossvalidation. Performance in fitting and predicting is demonstrated on some simple illustrative cases, and on three standard sets of real data: steroids binding to human corticosteroid binding globulin and testosterone binding globulin, and inhibitors of dihydrofolate reductase. © 1999 John Wiley & Sons, Inc. J Comput Chem 20: 1577–1585, 1999

Keywords: QSAR; quantitative structure–activity relations; partial least squares; Legendre polynomials; drug design

Introduction

Quantitative structure–activity relations (QSAR) studies have a long and diverse history, driven by the need to rationalize the observed biological activity of a training set of compounds in hopes of predicting the activity of a test set of compounds. In structure-based drug design, activity is known to involve binding to a particular site on a protein of known three-dimensional structure, so molecular modeling and specialized techniques can be used to dock small drug molecules to the site or suggest novel chemical structures having enhanced binding. Here, we focus on the more frequently occurring classical problem of correlating molecular structure to experimental binding data or even activity in cell

Correspondence to: G. M. Crippen; e-mail: gcrippen@umich.edu

Contract/grant sponsor: National Science Foundation; contract/grant number DBI-9614074

Contract/grant sponsor: The Vahlteich Research Award Fund

CCC 0192-8651 / 99 / 141577-09

and whole-animal assays, where even if the mechanism of action is known, the structure of the receptor is not. Although the underlying picture is still that of different small ligands interacting with some receptor pocket on a protein, few workers have tried to develop explicit site models,[1–5] and most have expressed their approaches in terms of the superposition of different molecules onto one particularly active one or onto the chemical structure common throughout the dataset.[6] Whether docking to a site model or superimposing molecules, there is an implied global search problem for the optimum positioning of each ligand that is surprisingly difficult, especially for conformationally flexible molecules.

A second disturbing trend is that the molecular features that are correlated with the observed activity tend to be rather specific, such as ''a carboxyl substituent at the 2-position,'' ''two hydrophobic groups separated by three rotatable bonds,'' or ''a pharmacophore consisting of a hydrogen bond donor 5.5 Å from a carbonyl group.'' It is difficult to go between precise and vague specifications of geometric and physico-chemical features in a smooth way. This leads in the first example to QSARs that are valid only for homologous series of compounds due to the implied but arbitrary superposition of a common ring structure such that common notational labels coincide in space for each molecule. In the other examples, the QSAR may be focussing on certain fine structural details that may explain the activity of a diverse training set but have limited predictive power because some other set of structural features is the real explanation.

Here, we present a set of molecular descriptors that range in a natural way from vague to precise, both in molecular geometry and in physico-chemical properties. Furthermore, these descriptors are independent of rigid translation and rotation (as topological descriptors are, for example), so that the issue of superposition or docking is avoided altogether. The descriptors also are independent of any ordering or labeling of the atoms or groups of atoms.

Methods

MOLECULAR REPRESENTATION

Each molecule is represented as a set of one or more conformation that are within 8 kcal/mol of the apparent global minimum, while each differs from the others by at least 1 Å in root-mean-square deviation of the nonhydrogen atoms after optimal superposition. Conformational searches were carried out with the random incremental pulse search implemented in the molecular modeling software, MOE,[7] employing the MMFF force field and its associated atomic partial charges. Each conformer treats the atoms as points having particular Cartesian coordinates, an atomic contribution to logP (octanol/water partition coefficient, ALOGP),[8] atomic contribution to the molar refractivity,[8] and empirical partial charge. Throughout this work we have consistently used these three properties, denoted by (h, r, q), respectively, but any other properties and any number of them could equally well have been used. The atomic properties are assumed to be independent of conformation but dependent on chemical structure and bonding.

LEGENDRE POLYNOMIALS

A convenient series expansion of a function defined over a finite interval is in terms of Legendre polynomials, $P_n(x)$, for $-1 \le x \le 1$ and $n = 0, 1, \ldots$ (see Table I). Most chemists have encountered them in the context of the radial part of the hydrogen atom wave function, but they are more generally used in all kinds of series approximations over finite intervals, much as Fourier series are used for describing periodic phenomena. Legendre polynomials are orthonormal in sense that

$$\int_{-1}^{1} P_n P_m \, dx = \delta_{n,m},$$

where the Kronecker $\delta_{n,m} = 0$ for integers $n \neq m$. Thus, an arbitrary function over the same interval can be approximated by

$$f(x) \approx \sum_i \left(\int_{-1}^{1} f(y) P_i(y) \, dy \right) P_i(x),$$

TABLE I.
Normalized Legendre Polynomials.

n	$P_n(x)$
0	$2^{-1/2}$
1	$(3/2)^{1/2}x$
2	$(5/8)^{1/2}(3x^2 - 1)$
3	$(7/8)^{1/2}(5x^3 - 3x)$
4	$(3\sqrt{2}/16)(35x^4 - 30x^2 + 3)$

where the approximation improves as more terms are added to the sum. Suppose $f(x) = 0$ everywhere except for n points $x_1, \ldots, x_n$. Then the coefficients in the expansion become sums

$$f(x) \approx \sum_{i=0}^{l} \left(\sum_{j=1}^{n} f(x_j) P_i(x_j) \right) P_i(x) \qquad (1)$$

and $f(x)$ is approximated up to level l. For example, suppose we want to approximate the distribution of atomic partial charges for a molecule having $n = 3$ atoms with $q_j = -0.8, +0.2, +0.5$. Here we are identifying the variable x_j in eq. (1) with the partial charges q_j, and setting $f(x_j) = 1$ for each of the three values because there is only one atom having the specified charge. Then the $i = 0$ term in eq. (1) is simply proportional to the number of atoms independent of their charges because $P_0(q)$ is a constant, independent of q. The next term represents the total net charge because $P_1(q)$ is a linear function of q. The next begins to show the differences between the negative and positive charges, and so on. As shown in Figure 1, the series approximation eventually approaches the given discontinuous distribution while keeping the total integral over the interval at 3 = the number of atoms. The central idea in VRI is that descriptors are related to distributions of certain atomic properties so that the effects of some arbitrary atom numbering or chemical nomenclature are removed. Then low-resolution descriptors are derived from the first coefficients of Legendre expansions of the distributions, and they amount to viewing a smoothed version of the original distribution. If greater detail is warranted, the coefficients from higher order terms can be included.

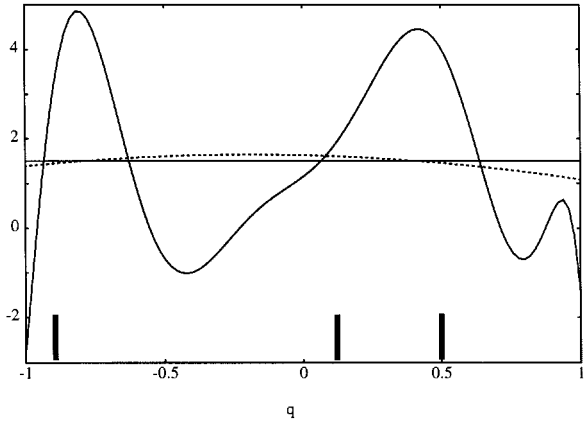

FIGURE 1. Legendre polynomial approximations to a charge distribution due to three atoms having q = −0.8, +0.2, and +0.5. The horizontal line is for level 0, the dotted line is level 2, and the complicated curve is level 9.

To describe a particular conformation of a given molecule, we use a distribution in four dimensions: the distance d_{ij} between each pair of atoms i and j, the total atomic hydrophobicity of that pair $= h_i + h_j$, the summed atomic contributions to molar refractivity $= r_i + r_j$, and the summed partial charges $= q_i + q_j$. Because Legendre polynomials work in the range $-1 \le x \le 1$, a linear scaling transformation is applied to each h_i, r_i, and q_i so that the extreme values for any atom in any molecule are mapped to -1 and 1. Atomic hydrophobicities range over all atom types in our scheme[8] from -3.1 to $+1.6$, so the sealed values are

$$h'_i = \frac{2(h_i + 3.1)}{1.6 + 3.1} - 1 \qquad (2)$$

Similarly, the empirical atomic contributions to molar refractivity assigned to atoms range from 0.8 to 13.8, resulting in scaling according to

$$r'_i = \frac{2(r_i - 0.8)}{13.8 - 0.8} - 1 \qquad (3)$$

Although partial charges for organic compounds are unlikely to exceed the range $[-1, 1]$, any value outside that range is converted to the limit

$$q'_i = \min[1, \max[-1, q_i]] \qquad (4)$$

For d_{ij}, one must choose some upper bound $M_d > d_{ij}$ for all molecules in the training set and likely test sets. Because this cannot be automatically determined just from the molecules in the current study, it remains the user's decision. Choosing M_d too large loses resolution by making most $d'_{ij} \approx -1$, while too small a value biases most of the scaled distances to $+1$, and may force assignment of yet larger distances to that limit.

$$d'_{ij} = (2 \min[M_d, d_{ij}] - M_d)/M_d \qquad (5)$$

Then in terms of the scaled quantities, the entire four-variable distribution is represented by the quadruple summation

$$f(d', h', r', q') \approx \sum_{k_d=0}^{l_d} \sum_{k_h=0}^{l_h} \sum_{k_r=0}^{l_r} \sum_{k_q=0}^{l_q} c_{k_d, k_h, k_r, k_q} \times P_{k_d}(d') P_{k_h}(h') P_{k_m}(r') P_{k_q}(q') \qquad (6)$$

where the coefficient c_{k_d, k_h, k_r, k_q} for a particular combination of levels of detail k_d, k_h, k_r, k_q in distance, hydrophobicity, molar refractivity, and partial charge is calculated as a sum over all pairs of atoms i and j:

$$c_{k_d, k_h, k_r, k_q} = \sum_{i \le j} P_{k_d}(d'_{ij}) \left(P_{k_h}(h'_i) + P_{k_h}(h'_j) \right) \times \left(P_{k_r}(r'_i) + P_{k_r}(r'_j) \right) \left(P_{k_q}(q'_i) + P_{k_q}(q'_j) \right) \quad (7)$$

Clearly, for an overall level of expansion $L = [l_d, l_h, l_r, l_q]$, there will be $(l_d + 1)(l_h + 1) \times (l_r + 1)(l_q + 1)$ terms, each with its corresponding coefficient. These coefficients constitute the majority of the molecular descriptors used for each conformation of each molecule. Note how all such descriptors are independent of atom numbering and any translation or rotation of the whole molecule, but they may depend on conformation.

Because even the [0, 0, 0, 0] descriptor is proportional to n^2 for a molecule having n atoms, we need an initial trivial descriptor having the constant value 1 for any molecule. Finally, eq. (7) is not only invariant under translation and rotation but also mirror reflection, so additional terms are needed to handle stereospecificity. Here we use a quantitative measure of chirality, $\chi(p)$ that depends on some atomic property p and the atomic coordinates of a given conformer.[9] This function always gives values of equal magnitude but opposite sign for the mirror image of a particular conformer. If mirror images cannot be distinguished because of symmetry, then $\chi = 0$; if mirror images are barely distinguishable due to similar but not identical atomic property values having nearly symmetrically related coordinates, then $|\chi|$ is small. Consequently, χ is meaningful in the present context as long as atomic position are distinguishable ($l_d > 0$) and some atomic property p is distinguishable (l_h, l_r, or $l_q > 0$). Thus, if $l_d = 0$, we append no further descriptors. Otherwise, we append $\chi(h)$, $\chi(r)$, and $\chi(q)$ when the corresponding levels are high enough. For example, for $L = [1, 1, 1, 1]$, there are $1 + 2^4 + 3 = 20$ descriptors.

DATA FITTING

In traditional QSAR we start with a matrix of descriptors $\boldsymbol{C}$ and a vector of observed activities $\boldsymbol{y}$, so that the mth row $\boldsymbol{c}_m$ of $\boldsymbol{C}$ is the set of descriptors for molecule m, and y_m is its observed activity. Then we seek a model vector $\boldsymbol{v}$ such that $\boldsymbol{Cv}$ approximates $\boldsymbol{y}$ in a least-squares sense, i.e., minimize $\|\boldsymbol{Cv} - \boldsymbol{y}\|^2$ with respect to $\boldsymbol{v}$. Here, the situation is different in two respects. First, each molecule is represented by at least one conformation, so for a flexible molecule m there are multiple rows of descriptors $\boldsymbol{c}_{ms}$ with conformers (structures) $s = 1, \ldots, 10$ typically. Secondly, in order to model both precise and imprecise observed activity, instead of a single value y_m we use a range $g_{l,m} < g_{u,m} \le 0$, where, for example, ''nanomolar to micromolar binding'' might be represented as $g_{l,m} = -9$ and $g_{u,m} = -6$; no observed binding would correspond to $g_{l,m} = g_{u,m} = 0$. If activity is enzyme inhibition, then the range $[g_{l,m}, g_{u,m}]$ corresponds to $\log K_i \pm$ error or $\Delta G_{\text{bind}} \pm$ error, so -9 means stronger binding than -6, for example. Then the calculated binding or activity of a conformer $g_{\text{calc},ms} = \boldsymbol{c}_{ms}\boldsymbol{v}$ is called superoptimal if $g_{\text{calc},ms} < g_{l,m}$, suboptimal if $g_{\text{calc},ms} > g_{u,m}$, and otherwise in-range. Then $\boldsymbol{v}$ is adjusted by minimizing the penalty function

$$\boldsymbol{F}(\boldsymbol{v}) = \sum_m \begin{cases} \displaystyle\sum_{\text{super } s} (\boldsymbol{c}_{ms}\boldsymbol{v} - g_{l,m})^2, & \text{any superoptimal} \\ \dfrac{n_m^2}{\left(\displaystyle\sum_{s=1}^{n_m} (\boldsymbol{c}_{ms}\boldsymbol{v} - g_{u,m})^{-1} \right)^2}, & \text{all suboptimal} \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

where clearly $F \ge 0$. In the case where all n_m conformers of molecule m are suboptimal, $F \to 0$ as any one or more $\boldsymbol{c}_{ms}\boldsymbol{v} \to g_{u,m}$.

Given how rapidly the number of descriptors rises with increasing levels of detail L, overfitting is a serious problem. As Wold et al.[10] have explained, a viable remedy is partial least squares (PLS), where $\boldsymbol{v} = \sum_{k=1}^{n_b} a_k \boldsymbol{b}_k$. The n_b PLS vectors $\boldsymbol{b}_1, \ldots$ are chosen to be few in number compared to the total number of descriptors, and they are chosen so as to capture as much as possible of the variance among each descriptor and their covariance with the observed values. Then eq. (8) is viewed as a function of the small number of adjustable parameters a_k. Of course, it is possible that an unfortunate choice of PLS vectors will exclude reaching $F = 0$, even though it might be possible using all the many components of $\boldsymbol{v}$ directly as independent variables.

ADAPTATION OF PLS

The original problem was the adjustment of some variable vector $\boldsymbol{v}$, given a descriptor matrix $\boldsymbol{C}$ and observed values $\boldsymbol{y}$, so that $\boldsymbol{Cv}$ approximated $\boldsymbol{y}$ in a least-squares sense. In standard QSAR, each row of $\boldsymbol{C}$ corresponds to a different molecule, and the columns correspond to the different descriptors, which depend on the chemical structure of the molecules, but not on the conformation. In this variant on PLS, each conformer of each molecule is viewed as a separate pseudomolecule, and the descriptors generally depend on conformation. For each row i we take $y_i = (g_{l,m} + g_{u,m})/2$, the mean observed activity of that molecule independent of conformation. This has the effect of saying there are possibly several pseudomolecules having the same observed activity, in spite of some variation in descriptors, so the analysis tends to emphasize common features of all conformers of each molecule, just as a conventional QSAR analysis tends to pick out common features of two different molecules that happen to have similar observed activity.

PLS works best on self-scaled data, so $\boldsymbol{y}$ and each column of $\boldsymbol{C}$ are independently rescaled to zero mean and unit variance. Denoting the mean and variance of $\boldsymbol{y}$ by μ_y and σ_y, and similarly the mean and variance of each column j of $\boldsymbol{C}$ by μ_j and σ_j, then the scaled activities and descriptors are given by

$$\begin{aligned} y'_i &= (y_i - \mu_y)/\sigma_y \\ c'_{ij} &= (c_{ij} - \mu_j)/\sigma_j \end{aligned} \tag{9}$$

Those columns of $\boldsymbol{C}$ that have no variance, such as the first descriptor being always 1, are deleted. The rows are divided into four groups that all span a full sampling of y_i values. Thus, if the rows are sorted by y_i values, the first group would be rows $1, 5, 9, \ldots$; the second would be $2, 6, 10, \ldots$; the third would be $3, 7, 11, \ldots$; and the fourth would be $4, 8, 12, \ldots$, etc. Then, as described in ref. 10, successive PLS vectors are added until a minimal prediction error is reached (PRESS = $\Sigma(\boldsymbol{Cv} - \boldsymbol{y})^2$ = prediction error summed over all four groups when for each group, the other three are used to determine $\boldsymbol{v}$). In a standard QSAR/PLS analysis this is simply a stopping heuristic to indicate that generating more PLS vectors is unlikely to enhance the predictive power of the model. Here it serves a similar function, but the connection to real predictive power is more remote for two reasons. First, for conformationally flexible molecules, any one of the four groups may contain more than one conformer of a given molecule, whereas a real crossvalidation of the entire model would involve removing all conformers of some molecules from the training set. Second, a standard QSAR predicts a single activity for each molecule, namely the components of the vector $\boldsymbol{Cv}$, whereas here each conformer s of molecule m has the predicted activity $\boldsymbol{c}_{ms}\boldsymbol{v}$, and the penalty function F in eq. (8) is satisfied if no conformers are superoptimal and not all are suboptimal, although some could be suboptimal by a wide margin.

In any event, once some number of PLS vectors are constructed, they are reconverted to correspond to the original unscaled $\boldsymbol{C}$ and $\boldsymbol{y}$. The kth scaled PLS vector is a linear combination of the scaled descriptors that is used to approximate the scaled activities $\boldsymbol{y}'$

$$\sum_{k=1}^{n_b} a_k \sum_j b'_{kj} c'_{ij} \approx y'_i \tag{10}$$

so that the corresponding unscaled PLS vector components $b_{kj} = b'_{kj}\sigma_y/\sigma_j$ approximate the unscaled activities

$$\sum_k a_k \left(\left(\mu_y - \sum_j b_{kj}\mu_j \right) + \sum_j b_{kj} c_{ij} \right) \approx y_i \tag{11}$$

which reintroduces the constant first descriptor

$$= \mu_y - \sum_j b_{kj}\mu_j.$$

For any other deleted column j of $\boldsymbol{C}$, the corresponding PLS vector component $b_{kj} = 0$.

TRAINING PROCEDURE

In summary, one selects a training set of compounds, performs a conformational search on each, determines a full set of descriptors for each conformer given a chosen detail level L, finds a minimal set of PLS vectors by pseudomolecule crossvalidation, and adjusts the a_k coefficients so as to minimize F in eq. (8). If this is unsuccessful ($F > 0$), the recommended procedure is to systematically raise the levels in L in order of increasing numbers of descriptors until the data can be fit. The final model consists of M_d, L, and $\boldsymbol{v}$. Given that we are fitting to observed ranges of binding, rather than a least-squares fit to single values, $\boldsymbol{v}$ is not uniquely determined. We carry out a simple

random search for any perturbation vectors w such that $F(v + w) = 0$, and include as many as 10 of these solutions as long as they differ sufficiently among themselves. Note that unlike least-squares fitting of a training set, each model either is in full agreement with the given binding intervals, or the model is rejected altogether.

PREDICTION

For subsequent prediction, the test molecules are prepared in the same way, and

$$g_{\mathrm{calc},\,m} = \min_s(c_{ms}v) \qquad (12)$$

for each model. This is an unambiguous prediction in that a particular model v applied to molecule m represented by a set of conformers $s = 1, \ldots, n_m$ produces a certain number, $g_{\mathrm{calc},\,m}$. However, it is possible for different models to produce similar satisfactory predicted activities for molecule m on the basis of different optimal conformers s, while for some other molecule, it might always be the same optimal conformer.

For either determining the models or predicting, the rate-limiting step is the conformational search for nonrigid molecules. Otherwise, CPU times are on the order of a few minutes on a Silicon Graphics workstation.

The many different statistics that have been used to express accuracy of fit to the training set and accuracy of prediction on the test set are not very well suited to multiple models fitting activity intervals. Here, we will simply use

$$E(\{v_k\}) = K^{-1} \sum_m \sum_{k=1}^{K} \max[0,\, g_{l,\,m} - g_{\mathrm{calc},\,m}(v_k),\; g_{\mathrm{calc},\,m}(v_k) - g_{u,\,m}]. \qquad (13)$$

Thus, if all molecules are predicted to be in-range by all K models, the error $E = 0$. Otherwise, the error is the mean over models of the total absolute value of deviations outside the observed activity range.

For all but the simplest examples, it is difficult to interpret a model in terms familiar to other QSAR methods. The calculated binding comes from a linear combination of many different geometric and physicochemical features, each of which is some sort of sum over the whole molecule. One cannot easily extract an explanation of activity from such a model in terms of substituent effects or pharmacophores. This is analogous to asking an X-ray crystallographer which atom in the unit cell is responsible for a particular diffraction spot. We hope to improve interpretation in future work. As it currently stands, the set of models derived from a training set of compounds can be used as a black box to predict the activity of any test compounds whatever, but inspection of the models themselves gives little direct guidance toward synthesizing new compounds.

Results

ARTIFICIAL EXAMPLES

Before considering standard test datasets of real experimental data, it helps to explain the method's performance in differentiating between simple pairs of molecules that differ in various ways. Suppose, for example, that CH_4 binds more weakly to some receptor than Cl_2 does. The simplest explanation is found at $L = [0, 0, 0, 0]$, where each molecule is described merely as [1, number of atoms pairs]. The coefficient of the constant term is negative, and that of the atom pair term is positive, so that methane is penalized relative to chlorine.

Consider the hypothetical case where 1,2-dichloronaphthalene and 1,2-dichloroanthracene both are active (large negative binding intervals), but the *meta* isomers of both bind poorly. Even though anthracene involves more atoms than naphthalene, and the *ortho* derivatives of both bind better than the *meta* derivatives, there is enough difference in the distribution of interatomic distances that at level [1, 0, 0, 0] the active compounds can be distinguished from the inactives. Only two PLS vectors were required for the four compounds, and the first distance-dependent descriptor is essential. Note that this model does not distinguish between atom types, because the h, r, and q levels are all zero.

To distinguish between the R and S isomers of CHFClBr, level [1, 0, 1, 0] triggers the calculation of χ for each, which gives nonzero values of opposite sign, because the four substituents are readily differentiable with respect to their atomic contribution to the molar refractivity. There are altogether $1 + 2^2 + 1 = 6$ descriptors, but because five of them are identical for the two molecules, the scaling procedure described in the Methods section automatically produces a model with nonzero terms only for the constant and χ descriptors.

Level [1, 0, 1, 0] also discriminates between *cis* and *trans*-1,2-dichloroethene, but for quite differ-

ent reasons. Because both molecules are planar, $\chi = 0$. The distance-dependent but type-independent descriptors are identical, as are the distance-independent but type-dependent ones. The final model involves only the constant descriptor and the one distance- and type-sensitive descriptor.

Finally, consider the R vs. S isomers of glyceraldehyde, which are so conformationally flexible they were represented by 20 substantially different low-energy conformers apiece. Of course, the distance-independent descriptors are the same for both compounds and were therefore not used. The simplest model required $L = [3, 0, 0, 2]$, which is a four-term expansion of the interatomic distance distribution coupled with a three-term expansion of the atomic partial charge distribution. Consequently, there is also one chirality descriptor that distinguishes atoms on the basis of their charges. The linear combinations of the 11 out of the 14 total descriptors used in the models were adjusted in terms of two PLS vectors. After the first model was determined, it was possible to find 10 variations on it, where sometimes individual terms vary in magnitude by 50% among the 11 solutions.

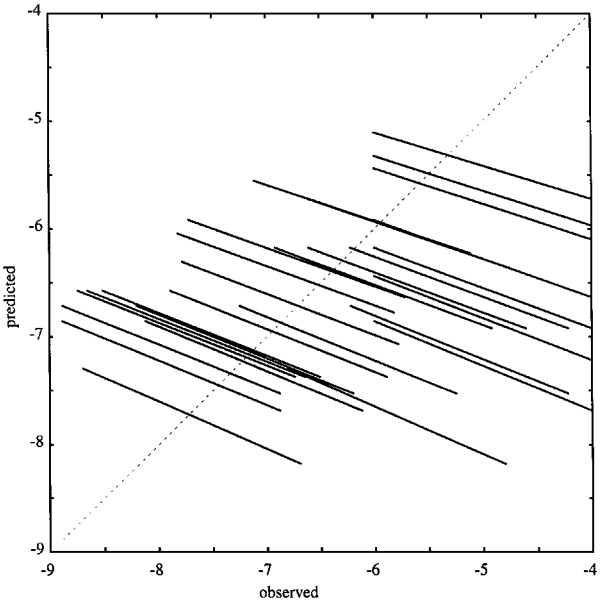

FIGURE 2. Observed vs. predicted binding of 28 CBG ligands by four site models. Each compound is represented by a solid line running from minimal observed and maximal calculated to the maximal observed and minimal calculated values. Any that cross the dotted line representing predicted = observed are at least not incorrect predictions.

BINDING TO CBG

A standard test case for many QSAR methods has been the binding of 31 different steroids to human corticosteroid binding globulin (CBG), denoted here as **1**–**31**. For the structures of these compounds, see Figure 5 of ref. 5. As in that study, we use for an observed binding interval the experimental values of the dissociation constant used in other tests of QSAR methods[4,11,12] transformed to $-\log K_{\text{diss}} \pm 1.0$. The assumed error limits are the same as in our previous work,[5] but do not necessarily correspond to the real experimental errors. An adequate training set, but not one with unique properties, is {**1**, **19**, **23**}. These three compounds are easily fit with level [0, 0, 0, 0], thus representing nothing more than the number of atoms in each molecule. Only one PLS vector is required, and four different models were found Of the 28 remaining compounds, 16 had predicted binding intervals extending no more than 0.1 outside the corresponding observed intervals; another seven extended less than 0.6 outside. Figure 2 shows the general agreement and the five compounds that were clearly incorrectly predicted by all four models, as indicated by five line segments that do not cross the dotted line. These five compounds are the only ones that contribute to a total prediction error [eq. (13)] of 5.78, i.e., 0.2 per compound on average over all 28 test compounds. By way of comparison, the Compass method of Jain et al.[12] used **1**–**21** as their training set, produced a single model that gives a single number for each test compound, and their standard deviation in predicting the $-\log K_{\text{diss}}$ for the remaining 10 compounds was 0.70.

BINDING TO TBG

A related standard test case is 21 of the same steroids, namely **1**–**21**, binding to testosterone binding globulin (TBG), which is usually thought to be more challenging than the CBG data. The observed dissociation constants are those used in previous studies[4,11,12] and are converted to binding intervals on a log scale as before. Small training sets can be fit with low levels of detail, but produce very poor predictions. If we start with only compound **1** and then successively add the worst predicted compound to the training set, eventually we reach a nine-compound training set, {**1**, **2**, **6**, **12**, **15**–**18**, **20**}, that can be fit at levels no simpler than [1,0, 2, 0]. We were able to find only one model, and seven PLS vectors were required

to reach that solution. As shown in Figure 3, six of the test compounds were completely correctly predicted. The total error $E = 1.43$ log units, i.e., a mean prediction error of 0.12 per compound. In comparison, other approaches[12] have had to use all 21 compounds in their training set, leaving none for prediction.

DHFR INHIBITORS

Another of our standard test suit are 48 inhibitors of dihydrofolate reductase (DHFR). These are conformationally flexible and chemically diverse, consisting of 23 4,6-diamino-1,2-dihydro-2,2-dimethyl-1-(substituted phenyl)-*S*-triazines,[13] 24 2,4-diamino-5-(substituted phenyl)pyrimidines,[14] and methotrexate, which is a pteridine derivative. See Tables 1 and 2 of ref. 5 for their chemical structures and observed binding constants. The experimental activity intervals we took to be $-\log K \pm 10\%$, except for methotrexate, which binds so tightly we simply used ''better than nanomolar,'' i.e., $[-12, -9]$. In attempting to train models, it soon became clear that both triazines and pyrimidines were needed in the training set, which was built up by successively adding the worst predicted compound, retraining, etc. Finally, for a training set of nine compounds consisting of triazines denoted as **1a**, **3a**, **4a**, and **5a** and pyrimidines **1b**, **2b**, **3b**, **20b**, and **24b** according to the labeling of ref. 5, we were able to find two models in terms of eight PLS vectors at level $[1, 1, 1, 1]$ that correctly predicted the binding of 11 compounds, namely **2a**, **8a**, **13a**, **14a**, **15a**, **23a**, **5b**, **6b**, **9b**, **13b**, and **15b**. In all, $E = 26.27$ or 0.67 per compound, and even methotrexate was underpredicted by only 0.4 (see Fig. 4). In comparison, our previous approach[5] was much more CPU intensive, required only six compounds in the training set, correctly predicted the binding of 19 compounds, had a lower value of $E = 10.9$, but methotrexate was underpredicted by 1.9 log units.

Conclusions

We have presented a new set of molecular descriptors that are independent of arbitrary translation and rotation of the molecule, as well as arbitrary atom labeling. Furthermore, these descriptors can be used to specify molecular features very vaguely or with gradually increasing detail. With adaptation of standard statistical methodology, one can fit a variety of artificial and real data sets involving chemically diverse compounds, stereoisomerism, and conformational flexibility. Because of differences in methodology, it is difficult to

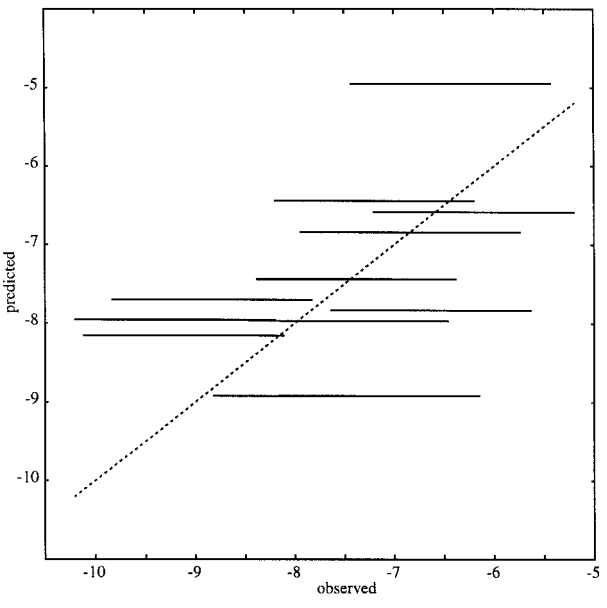

FIGURE 3. Observed vs. predicted binding of 12 TBG ligands by one site model, represented as in Figure 2. Because there is only one predicted value for each compound, the line segments are horizontal.

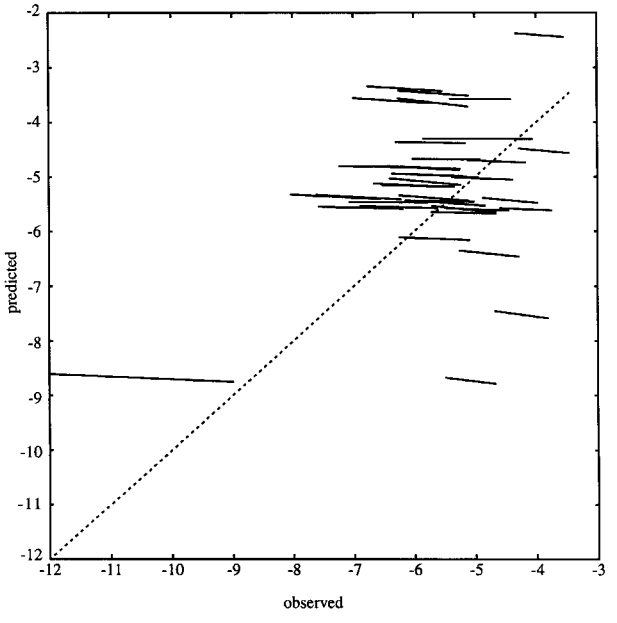

FIGURE 4. Observed vs. predicted binding of 39 DHFR inhibitors by two site models that differ only slightly in their predictions. The line segment at the lower left corresponds to methotrexate.

quantitatively compare with other methods the accuracy of fit to training compounds and the accuracy and statistical significance of the subsequent predictions. At least qualitatively speaking, the resulting models have predictive power that compares favorably with other methods, particularly because sometimes a smaller set of training compounds can be used.

Acknowledgment

The calculations in this work were performed using MOE molecular modeling software (Chemical Computing Group, Inc., Montreal).

References

1. Boulu, L. G.; Crippen, G. M. J Comput Chem 1989, 10, 673.
2. Bradley, M. P.; Crippen, G. M. J Med Chem 1993, 36, 3171.
3. Crippen, G. M. J Comput Chem 1995, 16, 486.
4. Schnitker, J.; Gopalaswamy, R.; Crippen, G. M. J Comp-Aided Mol Design 1997, 11, 93.
5. Crippen, G. M. J. J Med Chem 1997, 40, 3161.
6. Kubinyi, H., Ed. 3D QSAR in Drug Design; ESCOM: Leiden, 1993.
7. Chemical Computing Group, Inc., MOE version 1998.10, http://www.chemcomp.com..
8. Ghose, A. K.; Pritchett, A.; Crippen, G. M. J Comput Chem 1988, 9, 80.
9. Crippen, G. M. J Math Chem 1999, submitted.
10. Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., III. SIAM J Sci Stat Comput 1984, 5, 735.
11. Cramer, R. D. III; Paterson, D. E.; Bunce, J. D. J Am Chem Soc 1988, 110, 5959.
12. Jain, A. N.; Koile, K.; Chapman, D. J Med Chem 1994, 37, 2315.
13. Hansch, C.; Hathaway, B. A.; Guo, Z. R.; Dias Selassie, C.; Dietrich, S. W.; Blaney, J. M.; Langridge, R.; Volz, K. W.; Kaufman, B. T. J Med Chem 1984, 27, 129.
14. Hansch, C.; Li, R.-L., Blaney, J. M.; Langridge, R. J Med Chem 1982, 25, 777.